Introduction

Big Data Analytics

Presented by: Dr Sherin El Gokhy

Adv. Methods

Module 4 – Advanced Analytics - Theory and Methods

Module 4: Advanced Analytics – Theory and Methods

Part 7: Time Series Analysis

During this lesson the following topics are covered:

- Time Series Analysis and its applications in forecasting
- ARMA and ARIMA Models
- Reasons to Choose (+) and Cautions (-) with Time Series Analysis

Time Series Analysis

- Time Series: Ordered sequence of numerical values that measured in equally spaced values over time.
- **Time Series Modeling** deals with time based data. it involves working on time based data, to derive hidden insights to make informed decision making.
- Time Series Analysis is a method of prediction & forecasting
- **Time Series Analysis** is the analysis of sequential data across equally spaced units of time.

Time Series Analysis

- Time Series is a basic research methodology in which data for one or more variables are collected for many observations at different time periods.
- The time periods are usually regularly spaced and the observations may be either univariate or multivariate.
- **Univariate** time series are those where only one variable is measured over time, whereas multivariate time series are those, where multiple variables are measured simultaneously.

Time Series Analysis

The main objectives in Time Series Analysis are:

- To understand the underlying structure of the time series by breaking it down to its components.
  - Trend
  - Seasonality
  - Cycles
  - Random
- Fit a mathematical model and then proceed to forecast the future

Time Series Analysis

Trend component - Trend is a *long term movement* in a time series. It is the underlying direction *(upward or downward)* that can be *positive or negative depending on whether the time series exhibits an increasing long term pattern or a decreasing long term pattern.*

Seasonal component - It is the component of variation in a time series which is dependent on the time of the year. *It describes any regular variation with a period of less than one year.* For example, the average daily rainfall.

Cyclic component - Cyclical variations of non-seasonal nature, whose periodicity is un-known.

Random component - Random or chaotic values left over when other components of the series (trend, seasonal, and cyclical) have been accounted for.

Time Series Analysis

- Time Series: Ordered sequence of equally spaced values over time
  - **Time series** forecasting is the use of a model to predict future values based on previously observed values.
  - To forecast future events
    - Example: Based on sales history, what will next December sales be?
- **Method: Box-Jenkins (ARMA)**

Use Cases

Forecast:

- Next month's sales
- Tomorrow's stock price
- Time Series data provide useful information about systems generating the time series, such as:

Economics/ Finance: share prices, profits, imports, exports, stock exchange indices

Sociology: school enrollments, unemployment, crime rate

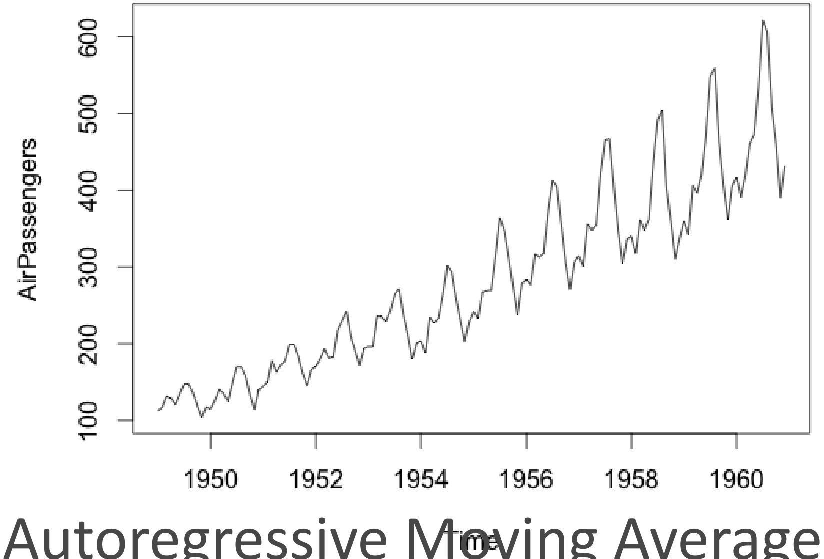
Environment: Amount of pollutants

Meteorology: Rainfall, temperature, wind speed

Medicine: Blood pressure measurements over time for evaluating drugs to control hypertension

Box-Jenkins Method: What is it?

- Models historical behavior to forecast the future



- Applies ARMA (Autoregressive Moving Averages)
- The **autoregressive model** specifies that the output variable depends linearly on its own previous values.
  - Input: Time Series
    - *Accounting for Trends and Seasonality components*
  - Output: Expected future value of the time series

Modeling a Time Series

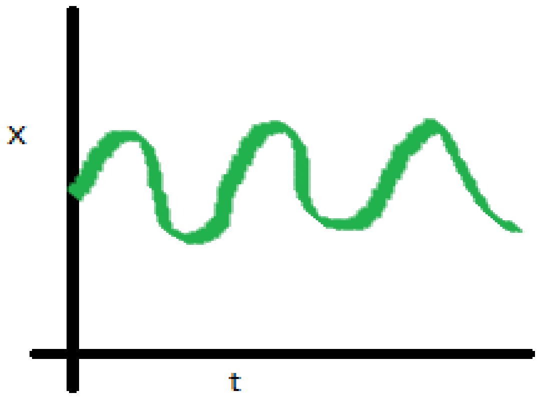
- Let's model the time series as

$$Y_t = T_t + S_t + R_t, \qquad t=1,\ldots,n.$$

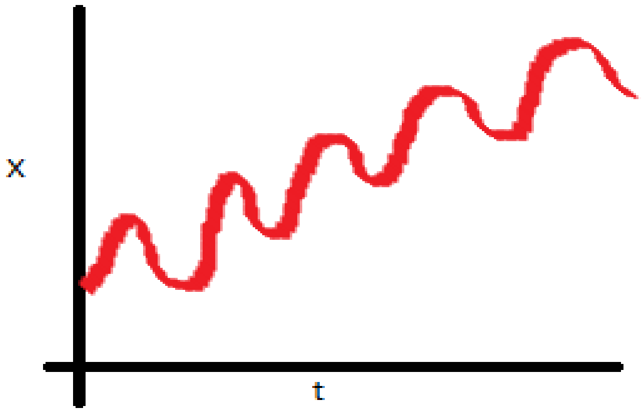
- T_t: Trend term
  - Air travel steadily increased over the last few years
- S_t: The seasonal term
  - Air travel variation in a regular pattern over the course of a year
- R_t: Random component
  - To be modeled with ARMA

Stationary Sequences

- Box-Jenkins methodology assumes the random component is a *stationary sequence*
- A stationary sequence is a random sequence in which the joint probability distribution does not vary over time. In other words the mean, variance and auto correlations do not change in the sequence over time.
  - Constant mean : the mean of the series should not be a function of time



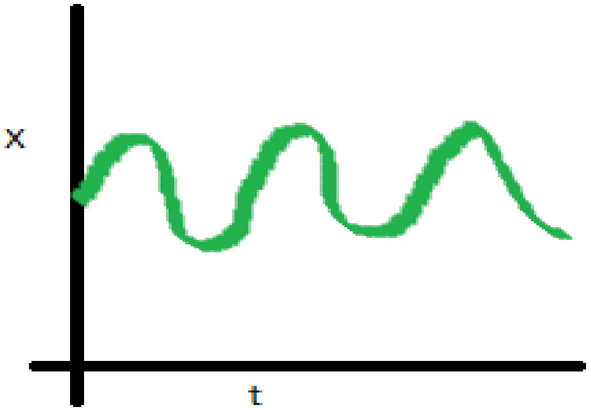
Stationary series



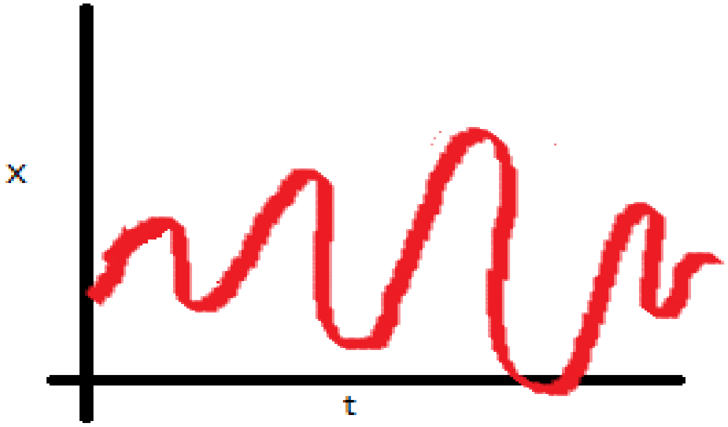
Non-Stationary series

Stationary Sequences

- Constant variance : the variance of the series should not be a function of time
  - The variance refers to the spread of the data set — how far apart the numbers are in relation to the mean



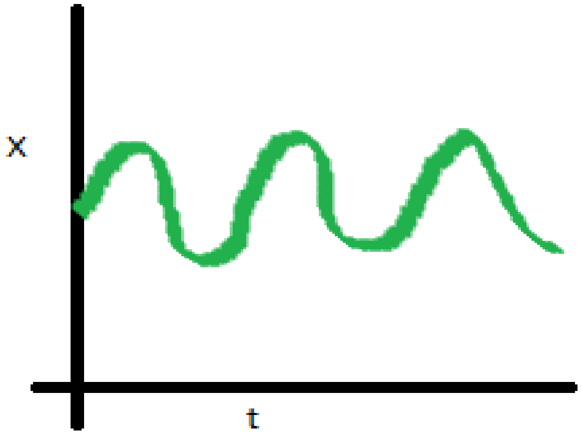
Stationary series



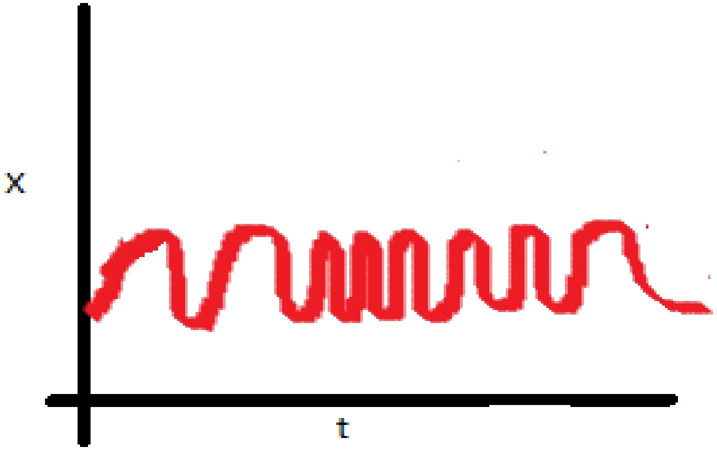
Non-Stationary series

Stationary Sequences

- Autocorrelation does not change over time
  - A **covariance** refers to the measure of how two random variables will change together and is used to calculate the correlation between variables.
  - Constant correlation of a variable with itself at different times



Stationary series



Non-Stationary series

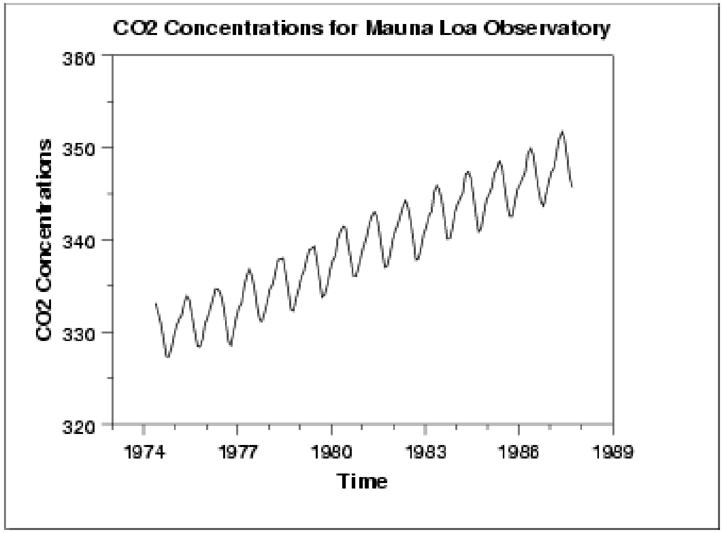
Stationary Sequences

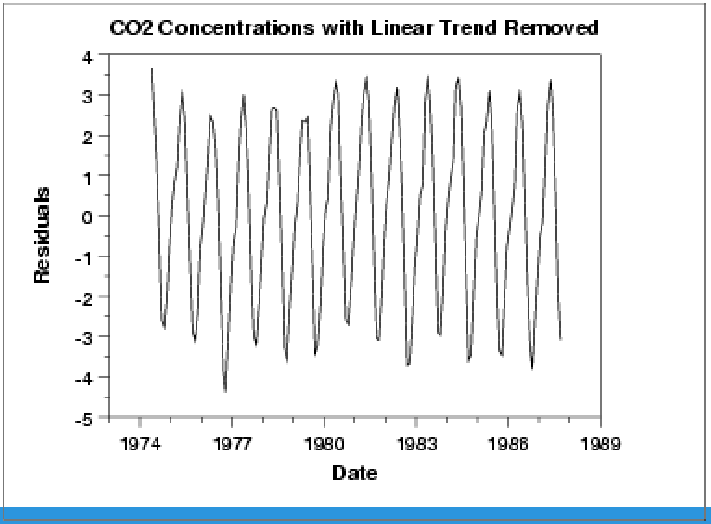
- Stationarity" implies that the series remains at a fairly constant level over time.
- If a trend exists (*the time series exhibits an increasing long term pattern or a decreasing long term pattern)*, then your data is NOT stationary.
- If your time series is non-stationary, you cannot build a time series model.
- In cases where the stationary criterion are violated, the first requisite becomes to stationarize the time series and then try stochastic models to predict this time series.
- In practice, to obtain a stationary sequence, the data must be:
  - De-trending
  - differencing
  - Seasonally adjusted

De-trending

De-trending is a pre-processing step to prepare time series for analysis by methods that assume stationarity

- In this example, the graphical plot of the data indicates nonstationarity, then you should "de-trend" the series.
- we see a linear trend, so we fit a linear model
  - $T_t = m \cdot t + b$
- The de-trended series is then
  - $Y^1_t = Y_t - T_t$
  - we simply remove the trend component from the time series.
- In some cases, may have to fit a non-linear model
  - Quadratic
  - Exponential





Differencing

- This is the commonly used technique to remove non-stationarity. Here we try to model the differences of the terms and not the actual term. For instance, **x(t) – x(t-1) = ARMA (p , q)**
- This is done by subtracting the observation in the current period from the previous one. If this transformation is done only once to a series, you say that the data has been "first differenced".
- This differencing is called as the Integration part in AR(I)MA. Now, we have three parameters **p : AR**

  d : I

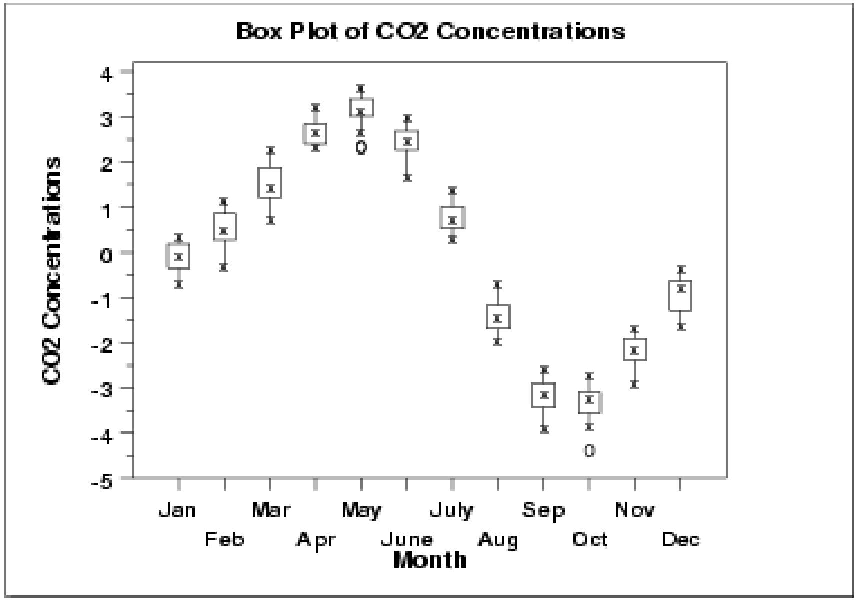
  q : MA
- This process essentially eliminates the trend if your series is growing at a fairly constant rate.

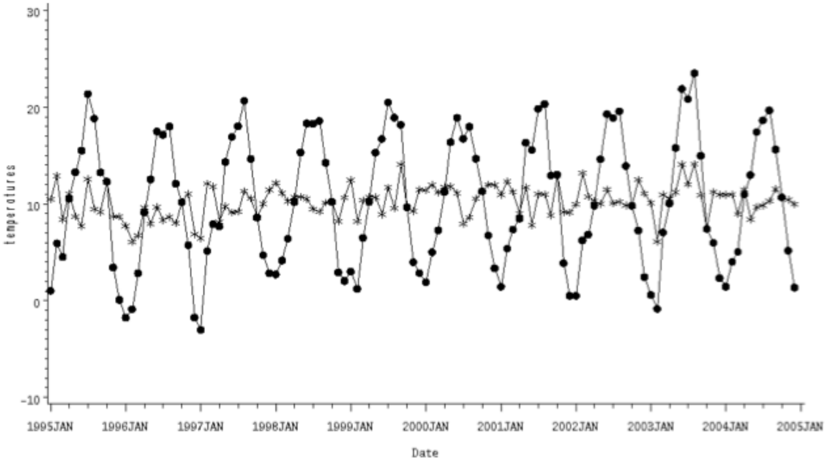
Seasonal Adjustment

- Plotting the de-trended series identifies seasons
  - For CO2 concentration, we can model the period as being a year, with variation at the month level

- A simple adjustment for seasonality is done with taking several years of data, calculating average value for each month and subtracting them from the actual value Y^1_t

$$Y^2_t = Y^1_t - S_t$$





Introduction to ARMA Time Series Modeling

- In ARMA model, AR stands for auto-regression and MA stands for moving average.
- Before we start, you should remember, AR or MA are not applicable on non-stationary series.
- In case you get a non stationary series, you first need to stationarize the series and then choose from the available time series models.
- AR model predicts Y_t as a linear combination of its last p values. An autoregressive model is simply a linear regression of the current value of the series on one or more prior values of the same series. The time series Y_t is called an autoregressive process of order p and is denoted as AR(p) process.

Introduction to ARMA Time Series Modeling

- A moving average (MA) model adds to Y_t the effects of a dampened white noise process over the last q steps.
- The simple moving average is one of the most basics of the forecasting methods. Moving backwards in time, minus 1, minus 2, minus 3 and so forth until we have n data points, divide the sum of those points by the number of data points n, and that gives you the forecast for the next period. So it's called a single moving average or simple moving average.

ARMA(p, q) Model

$$
\begin{aligned}
Y_t = & \boxed{\delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p}} \\
+ & \boxed{\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}}
\end{aligned}
$$

- The simplest Box-Jenkins Model
  - Y_t is de-trended and seasonally adjusted
- Combination of two process models
  - **Autoregressive**: Y_t is a linear combination of its last *p* values
  - **Moving average**: Y_t is a constant value plus the effects of a dampened white noise process over the last *q* time values (lags)
  - A **moving average** term in a time series model is a past error (multiplied by a coefficient)

ARIMA(p, d, q) Model

- ARIMA adds a differencing term, *d*, to the ARMA model
  - Autoregressive Integrated Moving Average
  - Includes the de-trending as part of the model
    - linear trend can be removed by *d*=1
    - quadratic trend by *d*=2
    - and so on for higher order trends

- The general non-seasonal model is known as ARIMA (*p, d, q*):
  - *p* is the number of autoregressive terms
  - *d* is the number of differences
  - *q* is the number of moving average terms

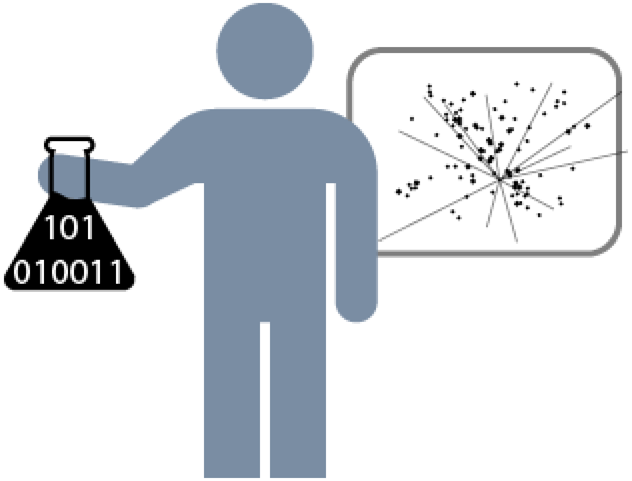
Diagnosis……ACF & PACF

- "Autocorrelations" are numerical values that indicate how a data series is related to itself over time. More precisely, it measures how strongly data values at a specified number of periods apart are correlated to each other over time.
- Auto Correlation Function (ACF)
  - Correlation of the values of the time series with itself
  - Helps to determine the order, q, of a MA model
- Partial Auto Correlation Function (PACF)
  - An autocorrelation calculated after removing the linear dependence of the previous terms
  - Helps to determine the order, *p*, of an AR model

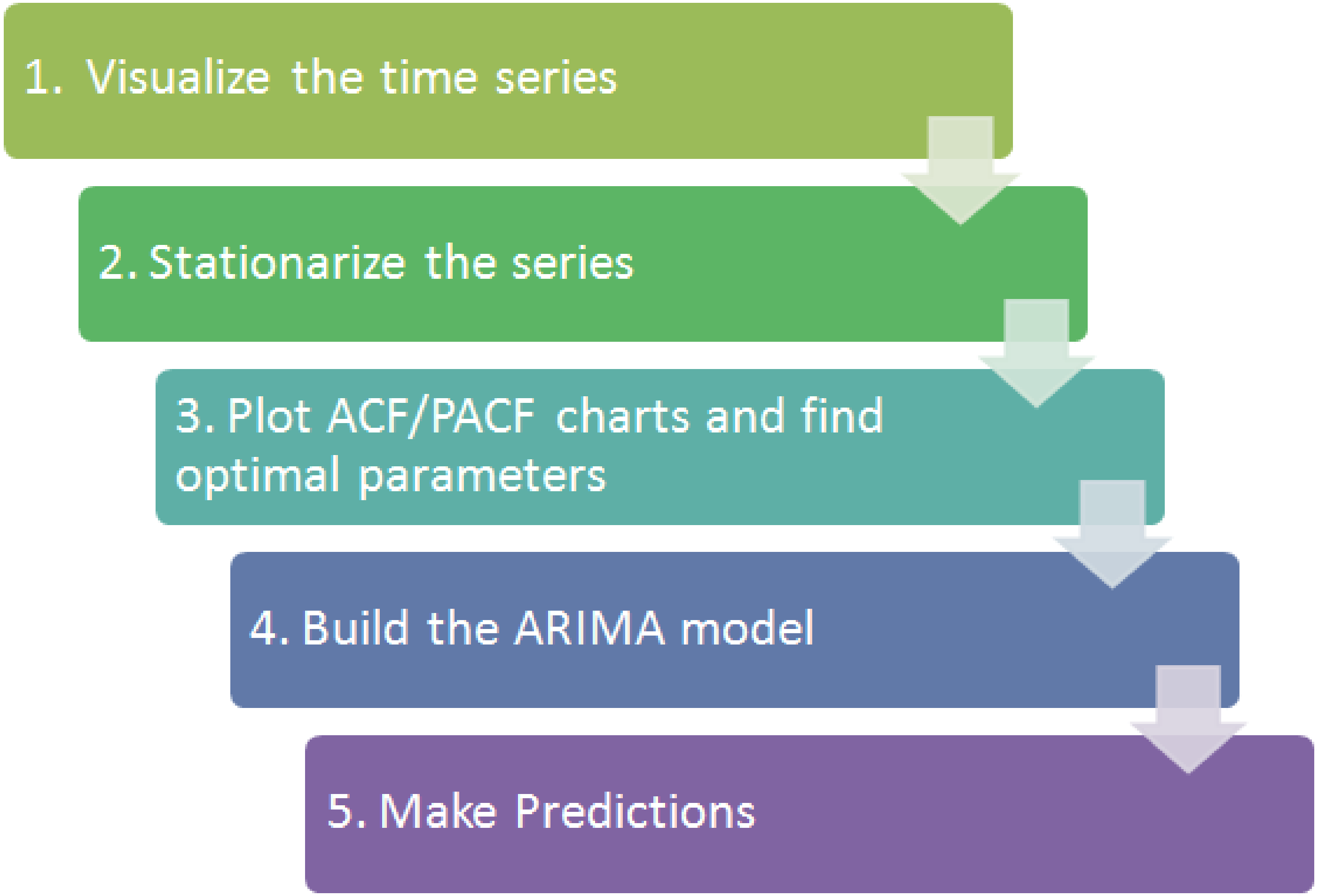
Model Selection

- Based on the data, the Data Scientist selects *p, d* and *q*
  - An "art form" that requires domain knowledge, modeling experience, and a few iterations
  - Use a simple model when possible
    - AR model (q = 0)
    - MA model (p = 0)

- Multiple models need to be built and compared
  - Using ACF and PACF



How to do a Time Series Analysis‘:



1. Visualize the time series
2. Stationarize the series
3. Plot ACF/PACF charts and find optimal parameters
4. Build the ARIMA model
5. Make Predictions

Step 1: Visualize the Time Series

- Exploring data becomes most important in a time series model – without this exploration, you will not know whether a series is stationary or not.

Step 2: Stationarize the Series

- Using the three commonly methods mentioned previously- de-trending, differencing and seasonal adjustment

Step 3: Plot ACF/PACF to find the optimal parameters

- The parameters p,d,q can be found using ACF and PACF plots. If both ACF and PACF decreases gradually, it indicates that we need to make the time series stationary and introduce a value to "d".

Step 4: Build ARIMA Model

- With the parameters in hand, we can now try to build ARIMA model. The value found in the previous section might be an approximate estimate and we need to explore more (p,d,q) combinations.

Step 5: Make Predictions

- Once we have the final ARIMA model, we are now ready to make predictions on the future time points. We can also visualize the trends to cross validate if the model works fine

Time Series Analysis - Reasons to Choose (+) & Cautions (-)

Reasons to Choose (+)	Cautions (-)
Minimal data collection Only have to collect the series itself Do not need to input drivers	No meaningful drivers: prediction based only on past performance No explanatory value Can't do "what-if" scenarios Can't stress test
Designed to handle the inherent autocorrelation of lagged time series	It's an "art form" to select appropriate parameters
Accounts for trends and seasonality	Only suitable for short term predictions

Time Series Analysis with R

- The function "*ts*" is used to create time series objects
  - **mydata<- ts(mydata,start=c(1999,1),frequency=12)**
- Visualize data
  - **plot(mydata)**
- De-trend using differencing
  - **diff(mydata)**
- Examine ACF and PACF
  - **acf(mydata)**: It computes and plots estimates of the autocorrelations
  - **pacf(mydata)**: It computes and plots estimates of the partial autocorrelations

Other Useful R Functions in Time Series Analysis

- **ar()**: Fit an autoregressive time series model to the data
- **arima()**: Fit an ARIMA model
- **predict()**: Makes predictions
  - *"predict"* is a generic function for predictions from the results of various model fitting functions. The function invokes particular methods which depend on the *class* of the first argument
- **arima.sim()**: Simulate a time series from an ARIMA model
- **decompose()**: Decompose a time series into seasonal, trend and irregular components using moving averages
  - Deals with additive or multiplicative seasonal component
- **stl():** Decompose a time series into seasonal, trend and irregular components using loess

Check Your Knowledge

Your Thoughts?

1. What is a time series and what are the key components of a time series?
2. How do we “de-trend” a time series data?
3. What makes data stationary?
4. How is seasonality removed from the data?
5. What are the modeling parameters in ARIMA?
6. How do you use ACF and PACF to determine the “stationarity” of time series data?

Module 4: Advanced Analytics – Theory and Methods

Part 7: Time Series Analysis - Summary

During this lesson the following topics were covered:

- Time Series Analysis and its applications in forecasting
- ARMA and ARIMA Models
- Reasons to Choose (+) and Cautions (-) with Time Series Analysis

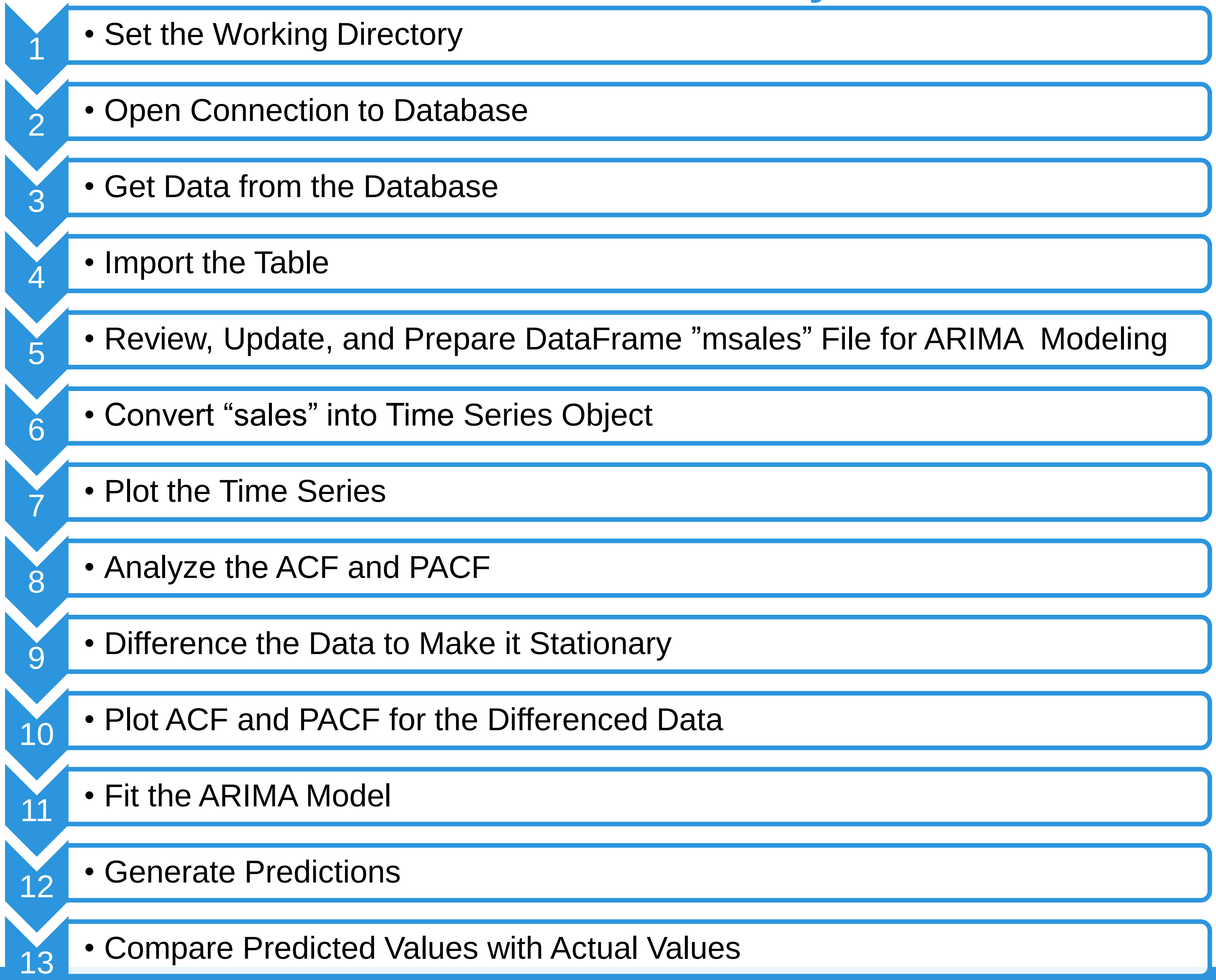
Lab Exercise 10: Time Series Analysis

This Lab is designed to investigate and practice Time Series Analysis with ARIMA models (Box-Jenkins-methodology).

After completing the tasks in this lab you should be able to:

- Use R functions for ARIMA models
- Apply the requirements for generating appropriate training data
- Validate the effectiveness of the ARIMA models

Lab Exercise 10: Time Series Analysis - Workflow



1. Set the Working Directory
2. Open Connection to Database
3. Get Data from the Database
4. Import the Table
5. Review, Update, and Prepare DataFrame ”msales” File for ARIMA Modeling
6. Convert “sales” into Time Series Object
7. Plot the Time Series
8. Analyze the ACF and PACF
9. Difference the Data to Make it Stationary
10. Plot ACF and PACF for the Differenced Data
11. Fit the ARIMA Model
12. Generate Predictions
13. Compare Predicted Values with Actual Values

Thanks